Interrogational Torture:
Or How Good Guys Get Bad Information with Ugly Methods

John W. Schiemann
Fairleigh Dickinson University

*Draft*
Please do not cite.

The author is grateful to Tobias Akerlund for research assistance and the figures and to Andrew Healy, Roger Koppl, Beer & Bull and Faculty Forum regulars at FDU, Marek Slaby, David Smailes, Georg Vanberg, Ken Benoit, Rene Lindstädt, Bruce Peabody, and three anonymous reviewers for their useful comments and suggestions.
Abstract

The use of “enhanced interrogation” techniques by the United States remains a matter of public debate. Recent revelations have revived the question as to whether harsh interrogation techniques amounting to torture such as waterboarding are an effective form of intelligence gathering. Answering this question is a necessary – if not necessarily sufficient – condition for the pragmatic justification of interrogational torture. Given the difficulty of approaching the question of effectiveness empirically, I address it theoretically, asking whether torture meets reasonable normative conditions of reliability, predictability, frequency, and intensity of torture. I find that although interrogational torture is neither reliable nor predictable it is likely to be used frequently and harshly once it is admitted as an interrogation technique.
The goal of *Perspectives on Politics* is “to provide political insight on important problems.” The use of “enhanced interrogation” techniques by the U.S. is an important legal and moral problem. It has preoccupied philosophers and legal theorists because, for many, the enhanced techniques amount to torture.\(^1\) To the degree that political philosophy is concerned with the proper balance between legitimate state authority, including violence, on the one hand, and individual autonomy, including autonomy of the body, on the other hand, then interrogational torture is also an important problem in normative political theory.\(^2\) This paper aims to provide insight on an important aspect of this problem made more salient by recent revelations such as the waterboarding of Khalid Sheik Mohammed and Abu Zubaydah 266 times: the effectiveness of interrogational torture in revealing clear and accurate information to the state. It may be that interrogational torture cannot be justified under any circumstances, but if it is to be justified at all, it must be effective. Interrogational torture’s effectiveness is a necessary, if not necessarily sufficient, condition for its justification.

Despite the importance of the question, there is little consensus among policy-makers, academics, or even interrogators on whether interrogational torture works. Members of the Bush administration and some interrogators have insisted that the techniques were valuable and prevented terrorist attacks.\(^3\) Other interrogators of high profile subjects have insisted such techniques did not work.\(^4\) Even under the Bush administration, the Department of Defense, the CIA, and the FBI differed as to whether torture produced valuable intelligence. And though the Obama administration has forsworn such methods, President Obama’s Director of National Intelligence, Dennis Blair, recently asserted that the techniques generated “high value information.”\(^5\)

Meanwhile, debate in academic circles has largely been confined to philosophy and to legal theory, with many citing interrogators in colonial Algeria, the Israeli security service, and the British in Northern Ireland to support their claims that interrogational torture is effective.\(^6\) They are joined by opponents of torture who, although ultimately arguing against it on other moral grounds, often assume that it works.\(^7\) Others in philosophy and legal theory have challenged the claims that torture is effective, often citing some of the same examples.\(^8\) The most comprehensive empirical study of torture and the
strongest case against assertions it is effective is made by Darius Rejali, who draws on extensive empirical research on conflicts around the world, including Algeria, the Gestapo, the Vietnam War, and police files to argue torture is ineffective. Nevertheless and despite the abundance of such evidence from around the world, Rejali acknowledges that “the empirical material...is...too fragmentary to allow for precise, validated causal claims.” As a result torture proponents are quick to dismiss such evidence as “a distracting and superficial numbers game.”

More recent efforts by the Senate Intelligence Committee to assess the effectiveness of the harsh interrogation program are unlikely to solve the numbers problem. Any empirical evaluation of observational data on interrogational torture is beset by what is the fundamental epistemological problem of interrogations generally: one can never know whether a detainee knew more information but did not reveal it. This is true whether or not the detainee was tortured and revealed information, was tortured and revealed no information, was not tortured and revealed information, or, finally, was not tortured and revealed no information. We do not – and will not – have this data, yet it is vital for determining whether torture is effective or not. There is the further, practical, problem in assessing any one interrogation method, torturous or otherwise, insofar as most intelligence comes from multiple sources and it is very difficult in reality to sort out the relative contribution of any one source. Finally, we can never know whether other, non-torturous methods would have revealed the same information.

In short, the data thus far made available, even that in Rejali, as well as any data likely to emerge, falls short of the social scientific standard required to draw firm conclusions, certainly for torture proponents. Yet determining whether interrogational torture works – defined as providing clear and valuable information to the state – is essential. First, if it does not work, then that means interrogators are not getting the information necessary to save lives. Second, if it does not work then the main rationale for its use – it is the only method we have in dire circumstances – disappears and any further use is pure sadism. Even (most of) its proponents would argue against it under these conditions. As Bagaric and Clarke admit in their utilitarian defense of torture, this is a “knock-down argument” for, “if this objection were valid [they] would change [their] minds and not countenance torture in any circumstances.”
This paper examines whether torture works well enough with a view to whether it should be state sanctioned, even under limited conditions and with restrictions such as the “torture warrants” suggested by Alan Dershowitz or the “torture thresholds” suggested by Bagaric and Clarke. By state-sanctioned I mean either explicit, legal permission or implicitly approved practice. On this definition, torturous interrogation techniques are (or at least were) state-sanctioned by U.S. officials. Given the absence of any reliable and systematic data from either the field or the laboratory on the effectiveness of torture, I investigate this question theoretically, the third (and worst) alternative of the three. Extending an insight by Roger Koppl in social epistemics, this paper employs the tools of game theory – the theory of interdependent, strategic choice – to examine whether torture is actually an effective strategy on the part of a state that generally prefers not to use torture, but will do so if it believes it necessary.

In particular, I examine the necessary and sufficient conditions for torture by a pragmatic – but not sadistic – state to generate equilibria – stable outcomes – that include clear and valuable information. Game theory allows us to build an analytical model of restricted, limited interrogational torture incorporating realistic elements and corresponding to what one might call the pragmatically normative view of interrogational torture – torture is unfortunately necessary in limited cases. It is then possible to compare the outcomes of the analytical model with the criteria defining the normative model of interrogational torture’s proponents. The analytical model’s outcomes then, substitute for the empirical data we would like to – but do not and will not – have to assess torture’s effectiveness. If the analytical model’s results satisfy the normative criteria, then the pragmatic argument for interrogational torture is supported; if the model outcomes fail to satisfy those criteria, support for even limited state-sanctioned torture falters. Since the goal is to address the problem of interrogational torture – not a problem in game theory – and since the intended audience includes all those interested in this problem – not just those interested in (or familiar with) game theory – the discussion in the main text is non-formal. The appendices provide the formal description of the model and proofs of the propositions.

A theoretical model cannot fully resolve the empirical question of whether interrogational torture is effective. But it is also true that it cannot be answered empirically. The modest goal of this paper is to
cast some light on an important problem with the tools available, recognizing that they are not ideally suited but also recognizing that the alternative is to ignore the problem entirely. It is also true that, given appropriately crafted assumptions, anything can occur in equilibrium. But competing models should be judged by 1) the verisimilitude of their assumptions to the world they model and 2) how much of the world is accounted for by the models’ results. Models should be judged in terms of how useful they are and it is possible to judge the relative value of competing models. The model that most closely reflects the empirical phenomena of concern in its assumptions and that accounts for more of the real world outcomes is superior to a model that either departs from those assumptions, fails to account for what we see empirically, or both. As I attempt to make clear below, I believe the model is defensible on both grounds and that makes its results useful.

The paper proceeds as follows. Section one outlines a normative model of pragmatic interrogational torture. The point of this model of torture-as-last-resort is to identify, first, constraints on the use of torture in interrogations and second, normatively acceptable consequences of interrogational torture in order to define how torture should work on the pragmatic view. Section two develops an analytical model of interrogational torture that is consistent with both the constraints of the normative model and other real world features of interrogations and torture in order to see how torture would work. Section three presents the results of the analytical model – the ‘data’ – and compares them with the predictions or acceptable outcomes of the normative model. The goal here is to see if the outcomes of the analytical model support or undermine the normative model. Section four summarizes the results and discusses their implications for state-sanctioned interrogational torture.

1. A Normative Model of Pragmatic Interrogational Torture

The pragmatically normative defense of limited, last-resort, interrogational torture rests on certain constraints on the way in which torture is employed as well as its acceptable effects on the information it generates. I focus on two sets of principles: the first set is normative principles concerning the regulation
of torture as state-sanctioned violence and the second set is epistemic principles concerning the information generated by torture.

The normative principles define two constraints on the use of torture by the state: frequency and intensity. If torture is used as a last resort by a democratic state on known guilty, knowledgeable detainees who fail to provide valuable information via other methods, then it should be used relatively infrequently. As a last resort, “torture should be confined to situations where the right to life is imperiled.”\(^{18}\) That torture should be infrequent makes more sense when one considers that torture is most effective as a threat whose very effectiveness is measured by the discontinuation (or even absence) of its use. I start talking to avoid future pain. Threats must be credible – it must be in the threatener’s interest to execute the threatened action – and some torture might be necessary to establish this credibility – but threats are usually costly to actually carry out. The point of the threat is to compel or deter particular behavior. A threat carried out has by definition failed to achieve its objective.\(^{19}\) Similarly, “when [torture] is used, the minimum degree of pain necessary should be used to obtain the information.”\(^{20}\) Whatever the merits of the legal reasoning behind them, recently released Bush Justice Department memos justifying interrogation methods amounting to torture clearly adhere to and emphasize the importance of these minimization criteria in their repeated references to safeguards, precautions, and escalation of techniques.\(^{21}\) Torture should be minimized in intensity as well as frequency.

In addition to constraints on the use of torture, the normative model of pragmatic interrogational torture also relies on two epistemic principles defining criteria for evaluating the consequences of interrogational torture: reliability and predictability.\(^{22}\) As Roger Koppl has shown, torture is an epistemic system, a social process designed to generate true judgments, and the measure of an epistemic system’s success is its “reliability: the ratio of true judgments to total judgments.”\(^{23}\) The implication here is that reliability means that the ratio of true to total judgments should be high. A reliable system produces many true judgments and not very many untrue judgments. Moreover, an effective method should be predictable; it should also be possible to predict when we should see those true judgments generated. In
other words, the system should produce a reasonably high probability of only accurate or truthful judgments. In the case of interrogational torture, this means clear and valuable information to the state.

These epistemic criteria have normative parallels in legal norms. Reliability, for example, is a bedrock principle in federal rules of evidence. The Daubert principle requires (among other things) the court to consider the “known or potential error rate” of some scientific procedure before it is accepted as scientific evidence. Presumably, this error rate should be not only known, but relatively low. Similarly, reliability and predictability are important for the justification of different forms of state-sanctioned violence. Police forces must undergo certain types of training in the use of deadly force and are governed by standard operating procedures. The 2008 U.S. Supreme Court decision upholding lethal injection as an instrument of capital punishment also illustrates. The plaintiffs argued that poor training meant that lethal injection is “sure or very likely to cause …needless suffering” and present a “substantial risk of serious harm,” or an “objectively intolerable risk of harm.” The Court upheld the method because they found that, in their view, the method if properly followed did not meet the identified thresholds; lethal injection is reliable and predictable enough to be constitutional. If it had not met these thresholds, presumably the court would have ruled the method unconstitutional. In other words, state-sanctioned violence should (nearly) always generate the desired results. In the case of capital punishment the desired result is a quick and painless death. In the case of interrogational torture, the desired result is clear and valuable information (reliability) and means that the procedure (nearly) always actually generates the clear and valuable information (predictability).

Finally, note that reliability and predictability are principles broader in applicability than rules of evidence or punishment and so are relevant to torture. As Robert Nozick demonstrates, “the right to be shown that [one] is being handled by a reliable…system” is an extension of the right to self-defense, and so is basic. In other words, this is not a particularly U.S. right, but rather a basic human right and so applies to non-citizens too.

In short, there are four normative criteria to assess the acceptability of the pragmatic account: the reliability and predictability of information and the frequency and intensity of the torture employed.
Given that we do not and will not have observational or experimental data to compare against these criteria, the remainder of this paper explores whether torture meets reasonable thresholds of the four criteria in a game theoretic model of interrogational torture. It will no doubt seem odd, even bizarre, to approach the question of torture with game theory. Upon reflection, however, it should not seem as strange as one might imagine. First, we do usually assume, albeit implicitly, that terrorist operatives are rational in some sense. Even those willing to die in suicide missions are unfortunately often all too effectively rational in carrying out their missions. Second, we also assume that interrogators are more or less rational. They have a clear goal in mind and they come up with very clever ways, not all of them torturous, to extract information from their detainees. Third, innocent detainees are no less strategically rational. Consider how Egyptian Abdallah Higazy reasoned after brought in for questioning about an aviation communications device found in his hotel room after 9-11 (it was later determined by the government not to be his): “If I say this device is mine, I’m screwed and my family is going to be safe. If I say this device is not mine, I’m screwed and my family’s in danger. And Agent Templeton made it quite clear that cooperate had to mean saying something else other than this device is not mine.”

Following this reasoning, Higazy falsely confessed. Fourth, the incentives created by an interrogation situation do, in fact, turn the interaction of a detainee and an interrogator into a kind of chess game in which each side attempts to figure out the other in an attempt to do what’s best for her. Interrogation is not fun and games, but it is a kind of game in the strategic sense. Indeed, Wantchekon and Healy have used a game of incomplete information to show how torture would emerge endogenously under reasonable conditions. This paper adopts a similar approach to a different end: whether torture is actually an effective information revelation technique.

Finally, and most importantly, game theory is actually well-suited to address the core criteria for deciding whether torture should be state sanctioned. In terms of reliability, a game theoretic approach permits identification of the necessary and sufficient conditions (if any) under which a detainee provides clear and valuable information in equilibrium. In terms of predictability, a game theoretic approach can identify whether the clear and valuable information outcome is unique and so predictable, or is only one
of two or more outcomes (other equilibria) and so is not predictable. Game theory also permits us to assess the frequency of torture by identifying how often torture – including unnecessary torture of an innocent detainee – occurs in equilibrium. Finally, it is possible to use some of the formal results from the model to derive some implications about the likely intensity of torture when it is used (comparative statics). In short, game theory permits the construction of an analytical model based on realistic assumptions that is consistent with a normative model of limited, pragmatic torture and compare the outcomes of the analytical model (equilibria, comparative statics) with the expectations of the normative model. The equilibria substitute for the empirical data we do not have but would use to assess the effectiveness of interrogational torture.

2. An Analytical Model of Pragmatic Interrogational Torture

The pragmatic interrogation game is composed of four basic elements: the players, their moves, the information they have when they move, and their payoffs to the different outcomes. Figure 1 illustrates the basic dynamic between the detainee and the interrogator.30

(Figure 1 about here)

Players. The game has two players, the detainee (D) and the state (S). In real interrogation settings both players will be uncertain about the other. The interrogator will usually be uncertain about whether the detainee is knowledgeable and weak, i.e. be willing to give up information in exchange for not being tortured, knowledgeable and strong, i.e. rather suffer torture than give up information, or innocent, i.e. not able to give any valuable information. This uncertainty is represented in figure one by the subscript i at the Detainee’s decision node indicating the detainee is one of three types – weak (D_w), strong (D_s), or innocent (D_z) – but the interrogator does not know which (the detainee knows its own type). The interrogator’s estimates about the likelihood the detainee is each type are p_w, p_s, and p_z, respectively.

Similarly, the detainee will be uncertain about the nature of the state’s interrogator – whether she is pragmatic (S_p), i.e. willing to use torture only as a last resort to extract information from a non-cooperative and knowledgeable detainee or sadistic (S_s), an interrogator that tortures even after
information has been obtained. While the state knows her own type, the detainee is unsure and thinks the state is pragmatic with probability $q$.32

**Moves.** The state can engage in two kinds of questioning: objective or leading. Under objective questioning, the state does not tell the detainee what she wants to hear. Under leading questioning, the state does let the detainee know what message would please the state. While leading questioning clearly provides no new or real information, the sad history of forced confessions from the church through the Salem witch trials to contemporary police and military practices reveals that some interrogators do engage in leading questioning. The Higazy case cited above is just a prominent example. This is important because in the first case an innocent detainee cannot provide the sought after information but in the second case, the detainee is able to tell the interrogator what she wants to hear. Thus, under objective questioning, the knowledgeable detainees (weak and strong) have two moves, $m$ and $n$, where $m$ is a message valuable to the state and $n$ is not, and the innocent detainee can send only message $n$. Thus $n$ is equivalent to keeping silent as well as providing information which is not valuable. This might mean information that is immediately perceived as non-valuable, or it could mean that the interrogators verified the information was non-valuable after some investigation. Under leading questioning, the innocent detainee can send message $m$ as well because the interrogator has revealed what she wants to hear, what she considers valuable, to the detainee. Al Qaeda operative Ibn al-Shaykh al-Libi, for example, told his Egyptian torturers in 2002 what their American counterparts wanted to hear: that there were close connections between Hussein’s Iraq and Al Qaeda. In 2004 he recanted (the CIA also withdrew the intelligence), saying he had asserted the connection only to avoid more torture.33 At the time any detainee makes his choice of $m$ or $n$ he does not know whether he is facing a pragmatic state interrogator or a sadistic interrogator for the state, captured by the subscript $j$ of the interrogator’s decision node, $S_j$. Following the detainee’s move, the state chooses to torture or not torture ($\tau$ or $\neg\tau$, respectively). Just as with the case of the detainee, neither state type knows, after receiving message $m$ or $n$, which type of detainee sent that message, captured by the subscript $i$ of the Detainee’s decision node, $D_i$. In other words,
she does not know whether she is questioning a weak detainee, a strong detainee, or an innocent detainee. The same is true for message \( n \).

**Information.** Before turning to how the players evaluate the outcomes, it is first important to note two important features of message \( m \) in the model. First, under both objective and leading questioning, the pragmatic state may believe a weak detainee sending message \( m \) actually has more information, that the detainee is still hiding something. If so, then the pragmatic state would want to torture to compel full disclosure. If not, if the detainee has revealed all he knows, the pragmatic state prefers not to torture. This uncertainty about information hiding is captured by a parameter \( \theta \), so that with probability \( \theta \) the weak detainee has told all he knows and with probability \( 1-\theta \) he is holding something back. Note that even if the model made the amount or degree of valuable information sent by the weak detainee the object of his strategic choice, this would not change the need to incorporate the pragmatic interrogator’s uncertainty about whether the detainee had still more information to divulge. Thus this set up captures the fundamental problem of interrogation – interrogator uncertainty about what the detainee knows – without unnecessarily complicating the model any further.

Second, under objective questioning, the pragmatic state does not always recognize a message \( m \) as valuable even when it is given. This could be because interrogators simply do “not believe the truths they hear.” For example, the torturers under the Pinochet regime in Chile, themselves highly religious, would not believe Sheila Cassidy when she implicated nuns and priests as collaborators with the opposition: “They found the truth more difficult to believe than the lies I had told them at first.” It might also be because the interrogators are unsure of what they seek. According to a French prisoner of the Gestapo, for example, “the Germans in many cases had no clear idea of what information they wanted and just tortured haphazard [sic].” Thus, the pragmatic state may recognize it as useful information (with probability \( \omega \)), but might also perceive it as false or otherwise not valuable information, equivalent to \( n \) (with probability \( 1-\omega \)). The detainee, however, is unaware of this uncertainty; a weak detainee sending message \( m \) assumes it will be recognized as valuable. Under leading questioning, this uncertainty does not exist; the state knows exactly what it wants to hear. In principle this uncertainty
about $m$ could be applied to the sadistic state as well, but since the sadistic state tortures regardless, the
degree of clarity of the message is irrelevant. The sadistic state tortures no matter what message the
detainee sends.

Making $\omega$ the private information of the state is an unorthodox assumption. The model adopts it
because this happens in the real world of interrogation. While it seems likely that most detainees will
know the state will suspect them of hiding information and that they cannot prove what they do not know,
they are likely to believe that the state will recognize valuable information if it is given to an interrogator.
Either it will be immediately seen as valuable or it can be verified later. Indeed, there is no point to torture
or any other form of interrogation unless the interrogator can recognize at least some information as
valuable. This is why the victim in the Chilean case above evinced surprise. In short, the model holds
empirical fidelity more dear than theoretical or methodological orthodoxy and “unfortunately, common
priors is a poor representation of the real world,” at least in some cases. What is an unorthodox
assumption from the perspective of game theory is a natural and necessary assumption from the
perspective of the real world of interrogation.

*Payoffs.* How the players evaluate the outcomes depend on each player’s type. Both the weak and
strong detainees pay costs $i$, for giving up valuable information (move $m$) and receive a payoff of 0 for $n$.
They also suffer physical and mental costs $k$, if they are tortured by the state and receive a payoff of 0 for
no torture. Both types of knowledgeable detainees (weak and strong) consider not giving up any
information and escaping torture as their ideal outcome, where there are no costs. The weak, however,
would rather give up information than suffer torture. The strong detainee prefers the pain of torture to
disclosing information. While the “simple folklore of pain” as in, for example, Bagaric and Clarke, has
given rise to the commonplace that ‘everyone talks, it’s just a question of when,’ the reality is that there
are many who do not break under torture, at least not in time to provide anything valuable to their
interrogators. Rumney points to court records of torture interrogations in France from the sixteenth to
the mid eighteenth centuries showing a failure to extract confessions ranged from a low of 67% to a high
of 95%. Rejali documents cases from around the world in which detainees endured horrific torture and
did not break. In the resistance to the Nazis across Europe, for example, “hardcore members did not normally break.”43 The CIA’s own 1983 interrogation manual asserts that “materialization of…fear is likely to come as a relief. The subject finds out that he can hold out and his resistance is strengthened.” “In fact, most people underestimate their capacity to withstand pain.”44 Both the weak and strong detainees consider giving up information and suffering torture to be the worst outcome possible, as in the Higazy case above.

The innocent detainee is identical to the weak detainee, except that, since the innocent detainee has no information to give, his cost to falsely confessing, λ, (sending message m under leading questioning) replaces the information costs i of the weak detainee.45 While it is theoretically possible for the innocent detainee to behave either like the strong or the weak detainee, depending on his aversion to lying, the model assumes that the innocent detainee mirrors the preference ordering of the weak detainee. Whereas the innocent detainee would prefer not to lie, to falsely confess to something he did not do, he is willing to do so if the alternative is telling the truth and getting tortured. The Higazy case is again an example.

Both states receive a payoff of V for detainee messages m that provide all the information they have to the state. Recall though, that there is some uncertainty about the value of message m. First, it may not include all the detainee knows. Second, the interrogator may not recognize it as valuable. For fractions less than full information, the states receive a payoff of V-δ. Since the value of m is private information on the part of the (knowledgeable) detainee, the interrogator’s subjective belief is that a message m provides V with probability θ and V-δ with probability 1-θ. As for the clarity of m, in the objective questioning variant of the model, the message m is perceived by the pragmatic interrogator as m with probability ω and is perceived as message n (a non-valuable message) with probability 1-ω. Recall that this uncertainty is the state’s private information; the detainee assumes that the state recognizes any message m as valuable (ω=1) and plays accordingly. The pragmatic interrogator assumes that the detainee treats ω just as the interrogator does and plays accordingly.46
Both state types pay a cost $\sigma$, if they fail to torture after receiving message $n$ from a knowledgeable. This parameter captures an element central to the rationale for using torture: establishing the credibility of the threat to use torture if the detainee does not provide valuable information. Moreover, it also captures the idea that states might (believe they) suffer reputation costs if detainees fail to provide useful information but the state still does not torture. An Al Qaeda training manual captured in Afghanistan provides support for this belief. According to an interrogator who read the manual, it viewed “America’s aversion to torture…a symbol of American weakness.”47 Failing to torture an innocent detainee who tells the truth (sends message $n$) nets both states a payoff of 0.

The two state types’ payoffs differ for using torture, however. Consistent with the normative model, the pragmatic interrogator prefers not to torture and will only do so as a ‘last resort.’ This is modeled by a cost $c$ for torturing any type of detainee. The costs $c$ might represent psychological or moral costs, reputation costs, morale costs, or any other cost. Moreover, since the pragmatic state uses torture only to extract information from knowledgeable detainees, she bears an additional cost $\alpha$, for “unnecessary” torture, i.e. torture of an innocent detainee who sends message $n$ (i.e. tells the truth) or of any detainee who reveals full information with message $m$). Adding the additional parameter $\alpha$ complicates the model but is necessary to be consistent with the normative model of pragmatic interrogational torture in which a pragmatic state prefers to torture only when absolutely necessary and does not want to torture unnecessarily. In contrast, the sadistic state receives a benefit $s$ to torture after any move by any detainee. Some states and/or some interrogators of some states (the two are identical from the detainee’s perspective) will torture whether or not they receive the information they want.

Before interpreting the model’s results in the next section, I would first like to further justify the specification choices in order to meet an obvious objection – craft the assumptions appropriately and anything can occur in equilibrium. First, the underlying model is a simple signaling game rather than a repeated game, which raises two potential problems. First, in a signaling game, what really counts is the signaling player’s underlying type, not the message she sends, whereas in this game, it is the signal that counts (a valuable message $m$ or a non-valuable message $n$), rather than the underlying detainee type.
Second, in this game the detainee does not move after the use of torture by the state, supposedly preventing the model from assessing the effects of torture on the detainee’s behavior. My response is a pragmatic one. It is sufficient to account for the behavior of interest while remaining as simple as possible. With respect to the first, however one conceives of the message vs. the player type, the players’ payoffs capture their preferences, preferences that are empirically realistic, and they act on those payoffs. That is all that is necessary for the model to be useful. Second, while it is true that torture is closer to what Schelling called a “compellance” rather than a deterrence game, it is nevertheless always the threat of (more) torture that in theory makes detainees give up information. The structure of the game captures this.

Another potential objection concerns the leading questioning variant of the model. Since there is no new information learned, only confession to what the interrogator wants to hear, this variant might seem to lie outside the bounds of interrogational torture, which is concerned with learning new (and valuable) information. Again, however, in the real world interrogators often, perhaps very often, engage in leading questioning, as repeated examples attest. The goal is to model what happens in an interrogation room in which torture is permitted; leading questioning happens and therefore must be modeled.

3. The Analytical and Normative Models Compared

This section discusses the results of the analytical model in terms of the epistemic and normative conditions necessary for torture to be state sanctioned. The following discussion ignores the behavior of the sadistic state, for two reasons. First, this state always prefers to torture, so there is nothing new learned. More importantly, the question at hand is whether torture is sufficiently effective – and so should be sanctioned – in a pragmatic state, a state which prefers to torture only knowledgeable detainees and only if they do not reveal valuable information. It is necessary to include the sadistic state in the analysis because (detainees will believe that) sadistic interrogators exist and this fact alters the behavior of detainees, but our interest is addressing the proper policy of the pragmatic state. Put another way, we know what a sadistic state actually does; what we want to know is what a pragmatic state should do.
The two variants of the model – objective questioning and leading questioning – together generate nine Perfect Bayesian equilibria in pure strategies. Three of the four equilibria under objective questioning, however, are behaviorally identical to their counterparts under leading questioning: they call for the same set of actions for each player. Moreover, interrogators in practice use both types of questions, so they are considered together. Collapsing these behaviorally equivalent outcomes together then, gives six equilibria, summarized in Figure 2.

(Figure 2 about here)

Equilibria are stable outcomes – combinations of player actions such as ‘give valuable information, use torture’ – in which no player wants to change her strategy given her beliefs and the beliefs are based on the actions of the players. Consider the two equilibria under leading questioning first. In the ambiguous information, no torture equilibrium, both the knowledgeable and weak detainee and the innocent detainee tell the interrogator what she wants to hear (message $m$), and the interrogator does not torture. The substantive interpretation of this equilibrium is that the interrogator does not know whether the information provided is objectively true or falsely confirmed by an innocent detainee trying to escape torture. The false confession, no torture equilibrium is similar, except that in this equilibrium only the innocent detainee sends message $m$, falsely confessing to what the interrogator wants to hear. A valuable information, surprise torture equilibrium occurs under objective questioning. In this equilibrium, a knowledgeable detainee provides information but is tortured anyway because the interrogator either doesn’t recognize the information as valuable or believes the detainee has more information to divulge. The remaining three equilibria occur under both objective and leading questioning. The valuable information, no torture equilibrium is the closest to the ideal outcome for the pragmatic defense of interrogational torture: a weak and knowledgeable detainee provides the interrogator with information and is not tortured. In the remaining two equilibria, no detainee provides information to the interrogator. In one case the state tortures because the interrogator believes the detainee is knowledgeable (the no information, torture equilibrium) and in the other the state does not torture because the interrogator believes the detainee is innocent (the no information, no torture equilibrium).
Before comparing the equilibria to the normative criteria it is worth noting that there is, in some sense, nothing new here. Each of the equilibria correspond both to intuition and to outcomes in the real world, outcomes identified both by defenders of interrogational torture (the valuable information, no torture equilibrium) as well as those identified by opponents of interrogational torture (the other five equilibria). The power and usefulness of modeling interrogational torture is not that it generates outcomes we do not see in the real world, but rather that it provides the underlying logic and coherence to the varied outcomes that we do see empirically. Had we either seen outcomes that do not occur empirically or had we failed to see important outcomes which do occur empirically, the model would be less powerful.

Because the equilibria are defined not just by the player actions but by their beliefs, each equilibrium can be thought of as inhabiting an area defined by the combination of the two players’ relevant beliefs. These beliefs are defined by three parameters in the model and so can be captured in a three-dimensional space, providing a more readily interpretable way to compare the analytical outcomes – substituting for data – with the predictions or criteria of the normative model. Since this comparison is made using this parameter space, it will be useful to describe the space in figure 3 now before turning to those comparisons.

(Figure 3 about here)

The vertical axis, labeled $q$, represents the weak and innocent detainees’ belief that the state is pragmatic rather than sadistic; the closer to the top (to one), the higher the likelihood the state is pragmatic and not sadistic in the estimation of the detainee. This belief is important because it determines whether weak and innocent detainees want to divulge information or not (send message $m$ or message $n$, assuming leading questioning for the weak detainee). Similarly, the horizontal axis labeled $p$ represents the pragmatic state’s belief that the detainee is innocent while the horizontal axis labeled $\theta$ represents the pragmatic state’s belief that, upon observing message $m$, the detainee has revealed all he knows. The closer to one on the $p$ axis, the higher the likelihood the detainee is innocent and non-knowledgeable in the estimation of the pragmatic state. The closer to one on the $\theta$ axis, the more likely it is that the detainee has told the state all he knows. These beliefs are also important because they determine whether or not the
pragmatic state prefers to torture (because the state wants to torture after message $m$ if it thinks the detainee is hiding more information and does not want to torture after message $n$ if the detainee is innocent since the state prefers not to torture innocent detainees).

The superscripted positions on each axis in the figures that follow (e.g. $p^*$) represent thresholds capturing this switch in strategies. The thresholds, derived from the proofs of the equilibria in the appendix, divide each axis into two or more parts. For values above the threshold, the players take one action; for values below the threshold, the players take a different action. For an analogue, say you are considering attending a ball game, but there is a threat of rain. Your personal threshold for determining whether or not you attend the game will factor in both the probability of rain as well as how much you want to see the game; for the same probability of rain, you might be more willing to attend a competitive game compared to one you think will be a blow out. Your threshold will be different for the two games. Just as your decision about whether to go to a ball game might depend on your threshold, so do the actions of the detainee and the interrogator depend on their thresholds. The combination of actions in each equilibrium are thus supported by values above or below these thresholds.

Now, if torture is effective, it must be both reliable and predictable. Moreover, since the point of torture by a pragmatic state is to threaten a weak and knowledgeable detainee into providing information, but not to torture unnecessarily (after valuable information from a weak and knowledgeable detainee or of an innocent detainee), torture should be used sparingly. Finally, since torture is viewed as a last resort, only that minimum degree necessary needed to elicit the information should be used. Each of these suggests a specific prediction about the equilibria emerging from the pragmatic interrogation game.

3.1 Reliability

Reliability in epistemic systems means that the method generates accurate information; it elicits truth. In other words, the method has a high “ratio of true judgments to total judgments.” In the present context, this means that torture should produce clear and valuable information – ideally without using torture afterwards. If torture is reliable enough to be state sanctioned, then equilibria with knowledgeable
detainees providing valuable information followed by no torture by the state should constitute a
significant portion of the total parameter space occupied by equilibria. More precisely,

Reliability: A truly valuable message \( m \) (clear and valuable information from a knowledgeable
detainee, not a false confession from an innocent detainee) should occur in equilibrium and for
wide range of parameter values and variants of the model so that it occupies a significant portion
of the total parameter space taken up by all equilibria. Call this the Torture Justification Region
(TJR).

Testing this prediction entails determining 1) whether an equilibrium corresponding to the TJR exists, 2)
the size of the TJR given its existence, 3) the portion of the parameter space taken up by all equilibria, and
finally 4) the ratio of the TJR to the space determined in (3).

There are two equilibria in which valuable information is provided to the state, the valuable
information, surprise torture equilibrium and the valuable information, no torture equilibrium, depicted
respectively in figure four by the light and darker shaded regions.

(Figure 4 about here)

In the lightly shaded surprise torture equilibrium the weak detainee sends message \( m \) because it is
confident enough the state is pragmatic (i.e. its belief \( q \) is above the threshold \( \tilde{q} \)) and is confident that the
interrogator will believe he has divulged everything he knows (i.e. its belief that the interrogator’s belief
\( \theta \) is above the threshold \( \theta^* \)) but is tortured anyway because the value of the message falls below the
interrogator’s actual information hiding threshold \( \hat{\theta} \). This means both that a detainee who has divulged
information is tortured anyway and that the state either fails to recognize valuable information as in fact
valuable or believes that the detainee possesses more valuable information. Since the left-hand boundary
of this equilibrium is defined by the detainee’s ‘version’ of the interrogator’s information hiding threshold
(\( \theta^* \)) and the right-hand boundary is defined by the interrogator’s version (\( \hat{\theta} \)), then the size of their
difference determines the ‘width’ of this equilibrium. Thus, although valuable information is provided,
the fact that the detainee is tortured anyway and that the state may not recognize it as valuable means that it falls outside the TJR.

The TJR is thus the darker shaded region constituted by the *valuable information, no torture equilibrium*. The question in terms of reliability is how likely this equilibrium is. What is the likely size of the space taken up by the TJR? To answer this question, consider the three thresholds that support this equilibrium, \( \hat{q} \), \( \hat{\theta} \), and \( p^* \), taking \( \hat{q} \) first. This is the weak detainee’s threshold for telling the truth. For a belief above this threshold (that the state is pragmatic and so will not torture if the detainee tells the interrogator the truth), the detainee provides useful and valuable information; below this threshold and the detainee does not. By intuition, as the value of the information grows, so does that threshold at which the detainee is willing to divulge that information. In other words, the weak detainee must be very, very confident the state will not torture after \( m \) to make giving up the information worthwhile, making \( \hat{q} \) high and thus pushing up the threshold at which the weak detainee is willing to talk. The result is that the underside of the TJR moves up. Conversely, as \( i \) goes down, the less valuable it is, the more willing is a weak detainee to send message \( m \). If information is not very valuable, then the threshold at which to give it away can be lower, increasing the volume of the TJR towards the bottom. Thus, a knowledgeable detainee is most willing to talk about the information that matters the least. All these considerations drive up the threshold for valuable information: a detainee would have to be very confident he is facing a pragmatic state before sending message \( m \) and not \( n \). The horizontal dashed line, therefore, is likely to be closer to the top than to the bottom, reducing the TJR volume. At the limit, as the weak detainee becomes certain that he faces a sadistic rather than a pragmatic interrogator, the TJR shrinks to a flat rectangular prism. Of course, as discussed further below, the harsher the torture the lower is the threshold, driving the lower boundary of the TJR toward the bottom and increasing its volume.

Consider now the state’s information hiding threshold, \( \hat{\theta} \), and its innocent detainee recognition threshold \( p^* \). Each threshold (independently) affects the interrogator’s decision whether or not to torture the detainee. If the state is confident enough (above \( \hat{\theta} \)) that the detainee does not have more information,
the pragmatic state does not torture following message \( m \). This is the fundamental epistemological problem in interrogations generally and alluded to above: a detainee – innocent or knowledgeable – can never prove what he does not know and a knowledgeable detainee will use the interrogator’s ignorance of the detainee’s “message space” (what the detainee knows) to reveal as little information as possible. But precisely because the interrogator knows this incentive to hide information, “the greater the [interrogator’s] ignorance of the message space, the less confidence he will have in the veracity of the [detainee’s message].”\(^{54}\) If this lack of confidence drops below \( \hat{\theta} \), the interrogator tortures.

The innocent detainee recognition threshold \( p^* \) works in the same way. This equilibrium requires a willingness on the part of the interrogator to torture after receiving message \( n \) from a detainee – this is the threat that compels message \( m \) from the knowledgeable detainee. But the interrogator is only willing to torture after \( n \) if she believes the detainee is not innocent, i.e. below the threshold \( p^* \). For beliefs above this threshold, she chooses not torture after \( n \) and the equilibrium collapses (because the knowledgeable detainee has an incentive to switch from \( m \) to \( n \) since he will not be tortured). In short, in addition to \( q > \hat{q} \) (the detainee’s threshold for divulging information), the existence of the TJR equilibrium depends on \( \theta > \hat{\theta} \) and \( p \leq p^* \), as figure four makes clear. To see how likely these conditions are to hold, consider both what happens to these thresholds as different parameters in the model change as well as their likely values empirically. Since these thresholds are largely the function of the same parameters (\( \sigma, c, \alpha, \omega \), and \( p_\cdot \)) they can be considered them together.

As reputation costs (\( \sigma \)) go up, so do both the information hiding and detainee recognition thresholds. The higher the perceived costs of failing to torturing a knowledgeable detainee who refuses to provide information, the more likely is the interrogator to employ torture. Note that this is true independent of the value of the suspected information possessed by the detainee. Even if the information value is low, this threshold can still be high if the reputation costs are high because reputation costs are independent of whatever information a detainee possesses. Empirically, these costs are likely to be high for any state that tortures. If torture is to work as a threat, it must be credible. To make it credible the
pragmatic state must torture if it does not receive valuable information. Failure to do so makes it less likely for future detainees to provide valuable information. This is why, for example, it will torture a strong detainee on whom it has no effect. Interrogators who rely on torture must actually use it if they fail to get the desired information.

Consistent with intuition, the higher the costs to torturing, whether the general costs $c$ or the extra cost $\alpha$ to “unnecessary” torture, the thresholds at which the interrogator is willing to forgo torture go down. Conversely, when these costs drop, when the state is less concerned about either cost of torture, both thresholds go up and torture is more likely. The litigation investment made by the U.S. to defend its ability to use torture, President Bush’s willingness to use a veto to ensure its availability to U.S. interrogators, the fact that Khalid Sheikh Mohammed was waterboarded 183 times and Abu Zubaydah was waterboarded 83 times, all indicate torture’s costs to be low, at least as viewed by the Bush administration. It is, of course, an open empirical question just how much individual interrogators care about unnecessary torture, but statements by President Bush and Vice President Cheney on down to actions by CIA interrogators in Afghanistan, Iraq, and Guantanamo Bay, coupled with the hundreds of detainees released from US custody in Guantanamo, all suggest this concern is not terribly high.

The final parameter for information hiding threshold is the clarity parameter $\omega$. As it increases, the value of the information is clearer to the interrogator and the threshold at which she does not want to employ torture comes down. As clarity drops (i.e. falls from 1 toward 0), the threshold increases. The more the interrogator believes the information divulged by the detainee is not valuable, the more likely the interrogator is to employ torture. Of course, this factor will vary widely across detainees and situations. Even if, however, one assumes maximum clarity so that $\omega=1$, the interaction of the other parameters as well as their empirically likely values suggest a very high threshold, moving the dotted line to the right, toward one, and shrinking the TJR volume. For empirically likely values on all four of the $\hat{\theta}$ parameters, the TJR dwindles to a thin vertical rectangular prism.
The final parameter for the innocent detainee recognition threshold $p^*$ is $p_s$, the belief that the detainee is strong. As $p_s$ increases, $p^*$ also increases and moves toward the back. As the interrogator becomes more confident that the detainee is knowledgeable and strong, the threshold at which she is willing to stop using torture goes higher. As the likelihood the detainee is knowledgeable and strong decreases, the threshold decreases.

In fact, it is possible to narrow the likely range of $p^*$ more firmly by considering the interrogator’s prior beliefs about whether the detainee is innocent or knowledgeable and, if the latter, whether he is strong or weak. Prior beliefs refer to the interrogator’s beliefs about $p_w$, $p_S$, and $p_Z$ going into the game, prior to any move by a detainee. The model makes no assumptions other than these estimates are between zero and one and together they add up to one. But how are interrogators in the real world most likely to order $p_w$, $p_S$, and $p_Z$? This is an empirical question the answer to which depends on particular circumstances, but there are some relevant general considerations. There is abundant evidence that the vast majority of detainees are either innocent or are very low level figures, suggesting that $D_Z$ is most likely and so $p_Z$ is the greatest probability.\textsuperscript{55} Nevertheless, given that interrogations are a “guilt-presumptive process,” it is unlikely for an interrogator to believe the detainee before them is innocent.\textsuperscript{56} In some cases this will be because they will have solid, independent information on the detainee. Even, however, when they do not, the operating assumption of an interrogator must be that there is something about a detainee which makes them a candidate for interrogation in the first place. Add to this the well-established propensity for confirmation bias and interrogators are likely to assume the detainee knows something rather than nothing.\textsuperscript{57} Given a knowledgeable detainee, the question is then whether he is weak or strong. The very availability of torture presumes that it can work, that it is possible to break people. Hence the commonplace, “everyone breaks, the only question is when.” Indeed, the decision to waterboard Khalid Sheik Muhammed for the 183\textsuperscript{rd} time after 182 previous sessions was presumably because interrogators believed it would compel him to reveal information he had theretofore kept from them. Thus an interrogator is more likely to believe that the detainee can be broken and made to talk
rather than believe either that he will hold out no matter what or that he is innocent. If so, then it is possible to make the following proposition:

**Proposition 1:** if the pragmatic interrogator believes it more likely the detainee she faces is weak and knowledgeable \((p_w)\) than both of the other two types \((p_s\) and \(p_z)\), then \(p^* < \hat{p}\) and the rear boundary of the TJR (on the \(p_z\) axis) shrinks to under one-half.\(^{58}\)

The substantive effect of this proposition is a reduction in reliability; the TJR shrinks toward the front of the parameter space to less than halfway on the \(p_z\) axis, reducing the size of this equilibrium and so the range of values for which it holds true. In short, the movement of \(\hat{\theta}\) to the right and \(p^*\) to the front results in a TJR that is a thin and shallow cube. Even if interrogators are successful in driving down the threshold \(\hat{q}\) by increasing the severity of torture, the TJR still takes up a relatively small portion of the total parameter space.

To assess interrogational torture’s reliability, to assess the ratio of acceptable outcomes to unacceptable outcomes, it is necessary to determine the portion of the total parameter space taken up by all equilibria. Portions of the parameter space in which nothing happens (no equilibria) should not ‘count against’ the TJR in determining the reliability ratio. This is determined by ‘filling in’ the parameter space with all six equilibria, depicted in figure five

(Figure 5 about here)

There are only two areas of the parameter space in which no equilibrium occurs, indicated by the two small, white, cubes just behind the TJR and surprise torture equilibria. The remainder of the space is taken up by at least one of the six equilibria, including the TJR (again the darkest shaded region). As figure five illustrates, if the TJR approaches the relative size argued for above, then, its reliability ratio is very small.

### 3.2 Predictability

We have just seen that torture can work, but not reliably so: the conditions under which valuable information occurs are empirically unlikely and constitute a low reliability ratio. Indeed, the ideal
outcome – valuable information from the weak detainee (message $m$), truth-telling by the innocent detainee (message $n$) and no torture after either – never occurs in equilibrium, let alone for a unique set of parameter values. Nevertheless, if the TJR – however unreliable – is unique for some combination of parameter values, then the case for state-sanctioned torture might be strengthened. We would know it works for that (rare) range of parameter values. Moreover, this ability to predict when torture works is also a fundamental epistemological and normative requirement. State-sanctioned torture must be not just reliable, but predictably so: if employed, it should regularly and predictably generate valuable information and only valuable information. More precisely,

**Predictability:** The reliable equilibrium just described should be unique; there should be no other equilibria for the same range of parameter values, in the same parameter space as the TJR.

For game theory to generate predictions, there must be no multiple equilibria, at least for the same set of beliefs. That is, the spaces defining different equilibria should not overlap each other so that more than one outcome is possible for a single combination of beliefs ($q$, $p$, and $\theta$).

Neither the no information, torture, nor the false confession, no torture equilibria overlap with the TJR, leaving the remaining two equilibria, ambiguous information, no torture and no information, no torture as possibilities. Whether or not they overlap depends on the relative orderings of the three thresholds on the $p_z$ axis: $\tilde{p}$, $\hat{p}$, and $p^*$. As figure five indicates, the rear boundary of the TJR is defined by $p^*$. The front edges of the ambiguous information, no torture and the no information, no torture equilibria are defined by $\tilde{p}$ and $\hat{p}$, respectively. Thus, the TJR equilibrium is unique only when $p^*$ is less than both $\tilde{p}$ and $\hat{p}$ because only then is there no overlap between them (i.e. for $p^* < \tilde{p} < \hat{p}$ and for $p^* < \hat{p} < \tilde{p}$). When $p^*$ is in the middle position there is overlap with one of the other equilibria (with the ambiguous information, no torture equilibrium for $\tilde{p} < p^* < \hat{p}$ and with the no information, no torture equilibrium for $\hat{p} < p^* < \tilde{p}$). When $p^*$ is greater than both thresholds, there is overlap with both of the other equilibria.
Recall that $p^* < \hat{p}$ when the interrogator believes that the detainee is most likely to be weak and knowledgeable (i.e. $p_w$ is greater than both $p_s$ and $p_z$). If this is, as argued above, empirically likely, then there is likely no multiple equilibria problem with the no information, no torture equilibrium. If it can be shown that the conditions supporting $p^* < \hat{p}$ are consistent with those supporting the TJR equilibrium, then there is no overlap with the ambiguous information, no torture equilibrium and the TJR is unique. To investigate this, first recall also that there are two versions of the TJR equilibrium, one under objective questioning and one under leading questioning. Since the TJR equilibrium requires valuable information from a knowledgeable detainee but no false confession by an innocent detainee (who thereby gets tortured), the leading questioning version of the TJR equilibrium requires the innocent detainee to send message n. The innocent detainee does this when his belief the interrogator is pragmatic drops below his threshold $q^*$. But the ambiguous information, no torture equilibrium requires the innocent detainee to send message m, which he does when his belief the interrogator is pragmatic meets or exceeds his threshold $q^*$. As a result there is no overlap or multiple equilibria problem for the leading questioning variant of the TJR for $p^* < \hat{p}$. This is not the case for the objective questioning version of the TJR:

**Proposition 2**: If the TJR exists under objective questioning, it is nowhere unique and shares its entire space with the ambiguous information equilibrium.

The proof of this proposition is in appendix C. The logic behind it is that, in order for the TJR to remain unique, the interrogator’s belief that the detainee has revealed all he knows (captured by parameter $\theta$) must be low, indeed, less than $\hat{p}$, which is less than one-half. If $\theta$ is greater than or equal to $\hat{p}$, the threshold $\tilde{p}$ drops to zero, making $\tilde{p} < p^*$ and pushing the front of the ambiguous information, no torture equilibrium all the way to the front of the $p_z$ axis so that it swallows the entire TJR. But a low $\theta$ does not support the TJR equilibrium, which depends on a high $\theta$ for the interrogator not to torture and so induce the weak detainee to send message m. Thus, while $\theta$ must be less than $\hat{p}$ to make the TJR unique, it must be greater than $\hat{p}$ in order to support that equilibrium. If then, $\theta$ is low, there is no TJR.
equilibrium in objective questioning. If θ is high enough to sustain the equilibrium, then the objective version of the TJR is nowhere unique, as illustrated in figure six.

(Figure 6 about here)

Once again the TJR (for objective questioning) is the darker shaded region. The lighter shaded region surrounding it is the ambiguous information, no torture equilibrium in which both the weak and innocent detainees send message m under leading questioning; in this equilibrium, the state receives message m but does not know whether it is true information sent by a weak detainee or simply a false confession by an innocent detainee telling the interrogator what she wants to hear.

For the values on the parameters supporting the objective questioning variant of the TJR equilibrium, both equilibria are possible. Nor would the most commonly used equilibrium refinement, the intuitive criterion, eliminate any of the equilibria and so generate uniqueness for the TJR.60 It is important to note that the TJR shares its parameter space with a minimum of one other equilibrium; if the assumption that \( p_w \) is higher than both \( p_s \) and \( p_z \) is relaxed, then it is possible for the no information, no torture equilibrium to overlap as well, resulting in two equilibria sharing some part of that space. In short, even if the (generally empirically unlikely) parameters supporting the TJR were known to hold, it would not be possible to predict that the outcome will be the one sought by the normative model (valuable information, no torture) unless interrogators engaged only in objective questioning. Interrogational torture as practiced empirically does not have predictable outcomes.

3.3 Frequency of Torture

Recall that the pragmatic state uses torture only to elicit clear and valuable information and it is always the threat of ex post pain which is supposed to induce cooperation ex ante. To this degree, torture is a threat that has failed if it has been carried out. Torture is used best when it is used least – even in a pragmatic state or by pragmatic interrogators. Of course, it is necessary on the pragmatic account to torture knowledgeable detainees who refuse to send message m, such as the strong detainee. This is “necessary” torture. These considerations suggest the following prediction.
Torture Frequency: There should be minimal unnecessary use of torture: little torture after: 1) message m by the weak detainee and 2) after message n by the innocent detainee. Torture should not dominate the parameter space; it should be restricted to a small part of this space.

Three of the six equilibria entail unnecessary torture. In the valuable information, surprise torture equilibrium a knowledgeable detainee is tortured after revealing information and an innocent detainee is tortured for telling the truth. In the valuable information, no torture and no information, torture equilibria the innocent detainee is tortured for telling the truth. Thus, the torture of an innocent detainee for telling the truth (sending message n) occurs in three of the six equilibria, one of which is the TJR equilibrium. Thus, even in this justification region an innocent detainee is tortured for telling the truth. Indeed, unless the pragmatic state is willing to torture an innocent detainee for telling the truth, the valuable information, no torture equilibrium collapses and the TJR disappears. Interrogational torture can work (unreliably and unpredictably), albeit at the price of torturing innocent detainees, leading to the following proposition:

Proposition 3: In order for interrogational torture to generate valuable information, innocent detainees must be tortured for telling the truth.

The proof of this proposition follows immediately from the conditions supporting the valuable information, no torture equilibria under both objective and leading questioning in appendix A. The basic logic is that unless the interrogator tortures after a message n, the weak detainee has an incentive not to reveal information and send message n. It is only the threat of torture after n which compels the weak detainee to tell the truth. In practice this means torturing an innocent for telling the truth.

Of the remaining three equilibria in which the innocent detainee is not tortured, he must lie (falsely confess and send message m) in two of them (ambiguous information and false confession under leading questioning) to avoid torture. Thus, in five of the six equilibria, the innocent detainee is either tortured or falsely confesses; there is only one equilibrium in which an innocent detainee tells the truth and is not tortured for doing so (no information, no torture). This, however, comes at the price of a knowledgeable detainee getting away without revealing information.
The portion of the parameter space occupied by torture depends on the relative ordering of the thresholds along the $p_2$ axis. Since each threshold represents the point at which the pragmatic state switches her strategy from torture to no torture because her belief that the detainee is innocent ($p_z$) exceeds the threshold, there is no torture to the right of the rightmost threshold. If $\tilde{p}$ is greater than both $\tilde{p}$ and $p^*$, then torture takes up less than half of the parameter space. If $\tilde{p}$ or $p^*$ is greater than $\hat{p}$, then torture can occupy more of the space. Retaining the assumption that $p^* < \hat{p}$, consider what would move $\tilde{p}$ higher, past $\hat{p}$ and toward one.61

First, as the likelihood the detainee has revealed all he knows ($\theta$) drops, the threshold after which the interrogator does not torture goes up. The more convinced an interrogator is that the detainee has more information, the more likely she is to torture. Similarly, the more likely is the probability that the detainee is weak and knowledgeable, the higher is the threshold. An interrogator who believes she is facing a weak and knowledgeable detainee rather than an innocent detainee is more likely to employ torture. The threshold also goes up as the general costs to torture ($c$) go down and as reputation costs ($\sigma$) approach the unnecessary costs of torture ($\alpha$). In accordance, with intuition then, as the lower are the costs to torture, the more willing is the interrogator to employ it.

Empirically, the extra costs $\alpha$ of unnecessary torture are unlikely to be much higher than the reputation costs $\sigma$ and the costs to using torture under any circumstances, $c$, are also likely to be low, as argued above. Also as argued above, the interrogator’s belief that the detainee has revealed all he knows, $\theta$, is likely to be low and her belief that the detainee is knowledgeable and will give in to torture (i.e. $p_w$) is high. All these considerations drive the threshold $\tilde{p}$ higher, increasing the space covered by unnecessary torture.

Moreover, as similarly argued above, the pragmatic state recognition threshold $\hat{q}$ is also likely to be high, so that torture occupies a large portion of the space from the bottom towards the top of the parameter space. There is torture for the entire range of $\theta$ because there is torture after $m$ for $\theta$ under $\hat{\theta}$ and torture of an innocent for $\theta$ above $\hat{\theta}$. 29
3.4 Intensity of Torture

The normative model additionally constrains the degree of torture employed. The level of severity of the torture employed should approximate just that level sufficient to compel divulging the information. If unnecessary torture is used more frequently than the criteria generated by the normative model, perhaps its severity nevertheless remains within those normative limits. This generates the following prediction.

**Torture Intensity:** *When torture is employed, its intensity will approximate the minimum degree necessary to compel valuable information.*

In order to evaluate this prediction, recall from the discussion of reliability above that the weak detainee’s information-revelation threshold \( \hat{q} \) is determined by the ratio of the value of information to the costs of torture. Assuming the pragmatic state believes she is facing a knowledgeable rather than innocent detainee, what incentive does this ratio give to the pragmatic interrogator? A failure to receive message \( m \) by the detainee might mean two (non-mutually exclusive) things. The detainee’s information might be very valuable and it would take a correspondingly large increase in torture costs \( k \) to compensate (and put the detainee over the threshold). Or it may be that the information is not all that valuable, but \( k \) is so low that the threshold is still high enough to prevent the detainee from revealing the information. In either case, the interrogator has an incentive to increase \( k \) as much as possible: to increase the intensity of torture so as to drive down the threshold and compel him to divulge information.\(^{62}\) Indeed, given that \( \hat{\theta} \) is likely to be high and \( p^* \) is less than one-half, the only way for the state to increase the reliability, to increase the volume of the TJR, is to increase \( k \).

Now, \( i \) and \( k \) are fairly idiosyncratic values that will differ greatly across detainees, even within each type. Some weak detainees will have very valuable information and some will have only minimally valuable information. The same is true with the aversion to lying for innocent detainees (\( \lambda \)). Some innocent detainees will be more resistant to confessing falsely than others. Now consider the costs of torture, \( k \). This will also be idiosyncratic; different detainees will have different pain thresholds. But
whereas the values of i and \( \lambda \) are out of the control of the state, the cost of k is less so. The state may not know the subjective value of k for any given technique exercised on any given detainee, but the state does know that escalating the torture technique will increase k for any given detainee. And increasing k lowers the threshold making the detainee more likely to send message m. What is the implication of the relationship between i and k on the one hand, and the strategic incentives of the interrogator, on the other?

**Proposition 4**: Once torture is admitted as an interrogation technique, the ‘calculus of pain’ and the strategic incentives facing the interrogator result in increasingly harsh forms of torture.

Any time the interrogator observes \( n \) and believes the detainee is knowledgeable, she has the incentive to believe k was just not high enough and so to ratchet up the pain. Or that the information is so valuable (i is so high) that k must be correspondingly high to elicit \( m \). This is even greater incentive to ratchet up torture. In short, we should see harsher and harsher forms of torture. Torture is not only frequent, but also vicious in the pragmatic state.

4. Conclusion

Interrogational torture’s effectiveness in generating clear and valuable information is a minimally necessary condition for its normative defense on pragmatic grounds. Given the absence of data to evaluate this effectiveness, this paper attempted an evaluation via the least preferable – but only – method available: theoretically in a formal model of pragmatic interrogation. It constructed an analytical model consistent with both the pragmatic defense of limited interrogational torture – last resort torture as defended by its proponents – as well as real world features of interrogations. This permitted substituting the outcomes of the analytical model for the missing data and compare them to the predictions and constraints of the normative model.

The normative model requires that interrogational torture be reliable, that it generates clear and valuable information. An outcome corresponding to this prediction in the analytical model – the TJR – is possible, but the conditions supporting it are empirically unlikely. They require a knowledgeable detainee to believe the promise he will not be tortured if he divulges information – a promise impossible to make
credible – and an interrogator to be very confident the detainee has told all he knows. Even so, proposition one demonstrates that the ratio of valuable information to all information is low. Interrogational torture is not reliable.

The normative model requires that interrogational torture be *predictable*, that it not only *can* generate clear and valuable information, but that it *will* do so under known conditions. The analytical model demonstrated via proposition two that the conditions known to generate the TJR under objective questioning also generate the ambiguous information equilibrium, which fully envelopes it, thus creating a multiple equilibrium problem. This multiple equilibrium – predictability problem can be avoided by engaging only in objective questioning, something that apparently does not happen in real world interrogations, given the CIA’s and later the military’s reliance on SERE methods originally used by the Chinese to extract forced (and false) confessions.63 Thus, interrogational torture is not predictable.

The normative model places constraints on the amount of torture permitted in terms of both *frequency* and *intensity*. There should be little unnecessary torture of a knowledgeable detainee who reveals information and an innocent detainee who tells the truth. The analytical model demonstrated that 1) knowledgeable detainees who provide information are nevertheless tortured for reasonable values of the relevant parameters, 2) that interrogational torture can only provide valuable information if the state is willing to torture innocent detainees for telling the truth (proposition 3), and 3) that torture takes up a good portion of the parameter space for empirically reasonable values on key parameters. Finally, as proposition four argued, the nature of the logic of torture as well as the incentives given to pragmatic interrogators means that torture will be harsh and not simply the minimum necessary. Interrogators will continue to use torture and to increase its intensity in an attempt to ensure the detainee’s threshold is low enough to make him talk. Torture, once it is admitted into the interrogation room, will be frequent and harsh.

Even in the hands of a pragmatic state, then, a state which prefers not to torture innocent detainees and does not wish to torture if it receives valuable information, torture is far more likely to produce no information or false information than it is likely to produce valuable information.
Furthermore, torture is not predictable. The same combination of beliefs that make valuable information from a knowledgeable detainee possible also make ambiguous information possible. Consequently, the state will often be unable to distinguish between useful and non-useful information even when it is communicated. Finally, the pragmatic state will torture – and torture harshly – much of the time. It will torture innocent detainees and will even torture detainees who have already provided it valuable information.

It will no doubt be tempting to policy makers favoring torture to dismiss the model as too abstract and too removed from “the real world” to be of any use or value. But the model incorporates the real world in both its assumptions and its outcomes. The fact that the latter are validly derived from the former only strengthens the argument. First, in terms of assumptions, the model builds in the realistic possibility that some detainees are weak while others are tough and can hold out and still others are innocent but swept up – and that it is the interrogator’s job to figure out which is the truth. The model also explicitly incorporates the fact that in the real world information is sometimes ambiguous and that knowledgeable detainees will try to avoid giving up all their information. Finally, the model does not assume all interrogators are sadists; rather it assumes a pragmatic state prefers not to use torture if at all possible. This not only corresponds both to the U.S.’s stated policy (at least under the Bush administration) as well as to a normatively pragmatic account of justifiable torture as in, e.g. Dershowitz or Bagaric and Clarke, but also provides us analytical leverage: strong conclusions – lots of harsh torture without much useful information by a pragmatic state – from weak assumptions – little preference for torture by a pragmatic state.64

Second, in terms of outcomes, the equilibria in the model cover the range of real world outcomes: valuable information but also false confessions, ambiguous information, and no valuable information at all. In other words, the model gathers a wide variety of observations from both the contemporary press as well as comprehensive empirical work such as Rejali’s under one unifying framework.65 The model accounts for these varying outcomes with a single underlying logic. What that underlying logic reveals – and what is perhaps a contribution of this approach – is that the results of interrogational torture we
observe in recent press revelations have little to do with torture in and of itself and much more to do with the strategic incentives that define the game played by detainee and interrogator. Formal models are “tools for probing reality.” The question as to whether – in reality – interrogational torture actually works, actually provides us with vital information we otherwise would not get is one of the pressing moral questions of our time. The debate over this question suggests that this reality needs probing and the probing offered here suggests we reconsider whether the United States should sanction interrogational torture.
Figure 1. Pragmatic Interrogation Game
### Interrogation Game Equilibria

<table>
<thead>
<tr>
<th>Objective</th>
<th>Leading</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Leading</td>
</tr>
<tr>
<td></td>
<td>Ambiguous Information, No Torture</td>
</tr>
<tr>
<td></td>
<td>False Confession, No Torture</td>
</tr>
<tr>
<td>Valuable Information, Surprise Torture</td>
<td>Valuable Information, No Torture [TJR]</td>
</tr>
<tr>
<td></td>
<td>No Information, Torture</td>
</tr>
<tr>
<td></td>
<td>No Information, No Torture</td>
</tr>
</tbody>
</table>

**Figure 2. Equilibria of Pragmatic Interrogation Game**
Figure 3. Parameter Space for Pragmatic Interrogation Game
Figure 4. Valuable Information Equilibria
Figure 5: Reliability of Interrogational Torture
Figure 6. Predictability of Interrogational Torture
References


Baze ET AL. v. Rees, Commissioner, Kentucky Department of Corrections, ET AL., No 07-5439, p.2 (The Supreme Court, Apr 16, 2008).


Kleinman, Steven D. 2007. Statement before the House of Representatives, Committee of the Judiciary, Subcommittee on Constitution, Civil Rights, and Civil Liberties. Hearing on Torture and the


[http://www.spectator.co.uk/archive/features/14208/is-torture-always-wrong.html?SelectedIssueDate=24%20September%202005](http://www.spectator.co.uk/archive/features/14208/is-torture-always-wrong.html?SelectedIssueDate=24%20September%202005) (accessed April 14, 2008).


Posner, Eric A. and Adrian Vermeule. Should Coercive Interrogation Be Legal?


Rumney, Philip N.S. 2006. The Torture Debate: Is Coercive Interrogation of Terrorist


### Notes

1 See, for example, the two collections by Levinson (2004) and Greenburg (2006).

2 While there is debate over what exactly constitutes torture, the U.S. has prosecuted foreign nationals and American military personnel for the use of techniques such as waterboarding because they constituted torture (Pincus 2006; Rejali 2007, 172). Thus I refer to these techniques as forms of interrogational torture.

3 Mackey and Miller 2004, 477.

5 For a detailed review of FBI – CIA – DOD differences on interrogation techniques, see the May 2008 report by the Office of the Inspector General of the US Department of Justice. For contrasting statements by the Directors of the CIA and the FBI, see Mazzetti (2008). For Blair’s statement see Blair (2009, 1).


8 See for example Brecher 2007 and Rumney 2006.


10 Rejali 2007, 7.

11 Bagaric and Clarke 2007, 58.

12 There are other motivations for, and effects of, torture, such as deterrence and social control, but in this paper I address only interrogational torture: torture designed to extract valuable information (Shue 2004, 53). See Wantchekon and Healy (1999) for a discussion of other motivations for torture.

13 Bagaric and Clarke 2007, 53.

14 Dershowitz 2003, Bagaric and Clarke 2007, 38-39. Note that this paper addresses the question of whether torture is sufficiently effective to be state sanctioned and does not address the abstract, casuistic, and moral question of whether torture can ever be justified. These two questions are often – unhelpfully – conflated in the torture debate.

15 Koppl 2006.


17 If we had good data, we could simply look directly at how torture does work.

18 Bagaric and Clarke 2007, 35.

19 To take a more prosaic example, the point of threatening no dessert to get your child to eat broccoli is not to deny dessert but to compel broccoli eating. The need to deny dessert means the threat failed.

20 Bagaric and Clarke 2007, 36; also Dershowitz 2004, 259.
See, for example, Bybee (2002).

Rumney also examines the effectiveness of torture from the perspective of predictability and reliability, but his use of these terms differs somewhat from the use here; more importantly, his approach is rooted in an “examination of the evidence” (2006, 481). Given the inevitably anecdotal nature of the evidence, this approach falls victim to the numbers game problem and so fails to convince torture proponents.

Koppl 2006, 91, 94.

Rumney 2006, 482.


Baze et al vs. Rees 2008, 2.

Nozick 1974, 102.

Higazy v. Millenium Hotel 2007, 8.

Wantchekon and Healy 1999.

Figure 1 represents the game under leading questioning. Under objective questioning, the innocent detainee does not have move $m$. This is explained below.

Thus, the pragmatic and sadistic interrogators in this game correspond to the professional and sadistic types of torturer in Wantchekon and Healy (1999, 601-602).

The model adopt the labels “state” and “S” to keep the notation clear; I do not claim entire governments or countries are sadistic, only that an individual interrogator working on behalf of a state may be.


Since under objective questioning, only a knowledgeable detainee can send message $m$, the interrogator would know she faces a knowledgeable detainee, but not know whether he is weak or strong.

The cases of Khalid Sheik Mohammed and Abu Zubaydah are good examples. Zubaydah provided valuable intelligence under non-coercive methods but was then waterboarded “at least 83 times in August 2002” because CIA interrogators believed he was hiding more information (Shane 2009). “Waterboarding
Used 266 Times on 2 Suspects,” *The New York Times*, accessed on May 21, 2009:


36 Rejali 2007, 489.


38 Rejali 2007, 496.

39 In formal terms, $\omega$ is not a common prior but is the private information of the pragmatic state.

40 Smith and Stam 2004, 786. Smith and Stam provide an extended defense of heterogenous priors in their original paper and their response to Fey and Ramsay 2006 in Smith and Stam 2006.


42 Rumney 2006, 491.

43 Rejali 2007, 496 and chapter 21 *passim*.


45 The model uses the different notation to indicate clearly that the innocent detainee has no useful information to give up until the interrogator tells him what she wants to hear and asks him to repeat it.

46 In formal terms, the interrogator assumes $\omega$ is common knowledge.

47 Mackey and Miller 2004, 180.


49 The beliefs supporting those actions differ somewhat.

50 This is an informal (and incomplete) statement of perfect Bayesian equilibrium. For complete but still accessible statements, see Morrow (1994) and Gibbons (1992).

51 This is closest to the ideal, but not the ideal, because in this equilibrium an innocent detainee is still tortured for telling the truth (sending message n). This is discussed in more detail below.

52 Koppl 2006, 91, 94.

53 The comparative statics results supporting the following propositions are given in Appendix B.

54 Koppl 2006, 99.
According to Brigadier General Janet Karpinski, former commander of Abu Ghraib, “generally 90 percent of the security detainees being held at Abu Ghraib were just innocent, had no information at all” (Banbury 2005). According to a study based on the Department of Defense’s own data, 55% of Guantanamo “detainees are not determined to have committed any hostile acts against the United States or its coalition allies” and “only 8% of the detainees were characterized as al Qaeda fighters.” Six in ten were held at Guantanamo because they were “associated with” groups the U.S. considered terrorist organizations. This included, for example, an assistant cook conscripted by the Taliban. The 86% of the detainees who were captured by Pakistani or anti-Taliban Afghan forces “were handed over to the United States at a time in which the United States offered large bounties for capture of suspected enemies” (Denbeaux and Denbeaux 2006, 2-3).

See, for example, an interview with an Army Captain for the exception apparently proving the rule (Oppel 2009).

Kassin and Gunjodsson 2004, 41.

See appendix C for the proof of this proposition.

The proof of proposition 1 in Appendix C showed that $\hat{p}$ is less than one-half. Thus, if $\hat{\theta}$ must be less than $\hat{p}$ for the TJR to be unique, then $\hat{\theta}$ must be less than one-half.

Loosely speaking, the intuitive criterion rules out implausible beliefs off the equilibrium path which might support an equilibrium, in particular the belief a player would make any move the payoff to which is always lower than the payoff to some other move. See Gibbons (1992, 233-244) for a formal definition and explanation.

The conditions for $p^* < \hat{p} < \bar{p}$ are identified in the proof of proposition 2 in appendix C.

See comparative statics behind this assertion in appendix B.

See Mayer (2008, 164 and passim).

Rejali 2007. See Clarke and Primo 2008, 742 on the utility of models as measured by their ability to unify disparate empirical phenomena under one framework.

Brady 2004, 297.